FAKE NEWS DETECTION USING MACHINE LEARNING

NAME:HIBA NAFEESATH

ROLL.NO:20

PRODUCT OWNER:SYED FEROZE AHAMED M

Table Of Contents:

1. Description
2. Modules
3.Methodology
4.Future Enhancement
5. Developing Environment
6.Project plan
7.User story
8. Product backlog
9. Sprint plan
10. Sprint Actual
11.Technique-Natural Language Processing
12.Classifier
13.Algorithm
14.Conclusion

DESCRIPTION:

In traditional news making procedures, very limited and authorized individuals are involved and newspapers,radio,television were the only source of news. Due to these reasons news, credibility and authenticity are preserved. But in the era of internet, social network is becoming a news source of news. Easy and free access to these social networks makes the task of fabricating fake news and manipulating news a very effortless task. There is no authorize control point of these manipulated fake news which creates a question over there credibility and authenticity. The ease of getting direct news from the platform they mostly use has attracted the user. The reason to spread fake news can be social, political, and economical.

The rate at which it spread is very fast due to which controlling the spread manually is not possible.There is no platform via which the user can check the credibility and authenticity of the news and where authorities can directly inform about the fake news prevailing. Due to which people can believe in the news which can be a trouble for them and as well for society also. In the existing system, the action is taken after the adverse impact had already hit society. The proposed platform is useful for both common people and official authorities to prevent the spread of rumours in form of news. There is no existing platform that can verify the news and categorize it. This paper proposes a system that can be used for prediction of news to be real or fake. This system is based on natural language processing to extract features from the data and then these features are used for the training of machine learning

This system is to detect whether the news is fake or real by using Cosine similarity Algorithm in machine learning.when we give news it will take words,check frequencies and similarities with the model ,based on this the system will detect whether the news is fake or not.

MODULES:

1.Data collection

2.Data Cleaning

3.Data Preprocessing

4.Training

5.Testing

6.Result Generation

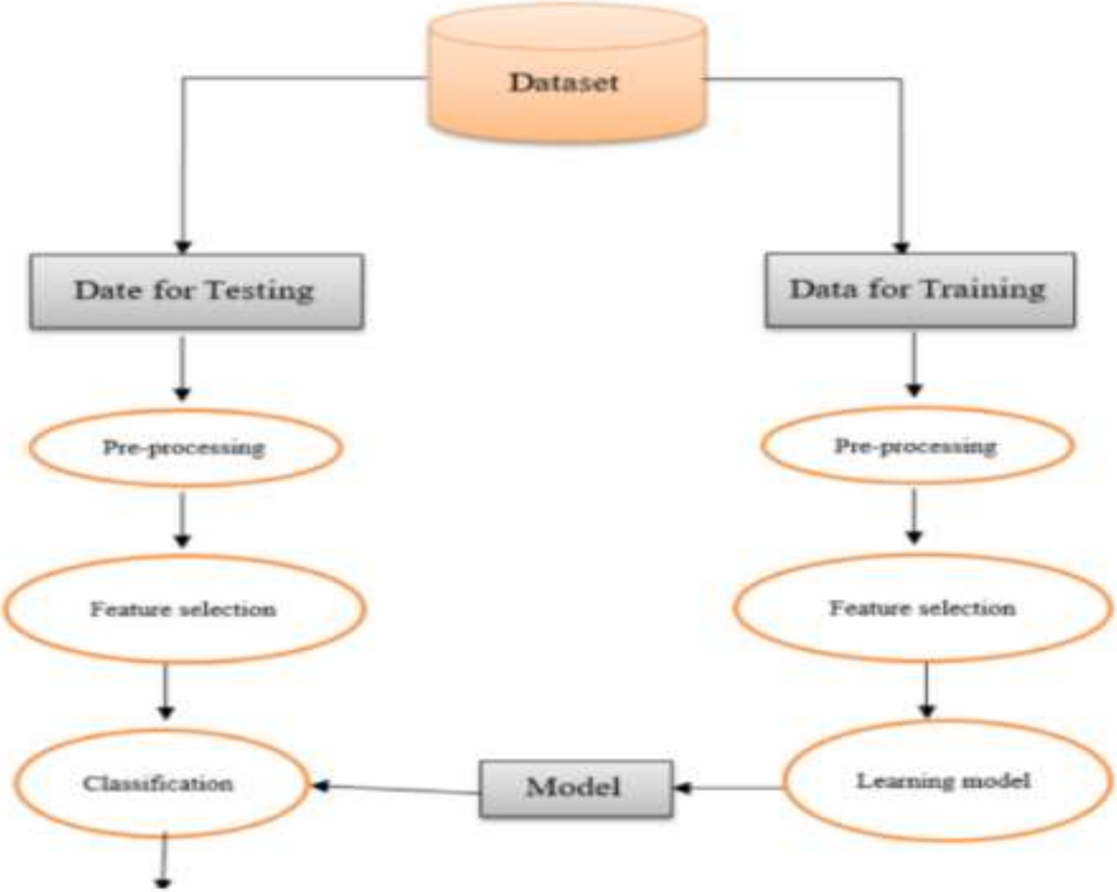

Figure 1. Describes the Proposed System Methodology

METHODOLOGY

Our objective is to find the best Machine Learning Algorithms and Natural Learning Process methods for the prediction of news. Then the best performing model will be saved and will be linked to a user interface by which it can predict new input data . To achieve this objective training data have to go through various intermediate processes before giving it to Algorithms. The main parameters on which the performance of the model will be judge are accuracy of the model.

1. Data

Data is the prime ingredient of this project, as these data features are extracted using Natural Language Processing. By using these features of the data, Machine Learning Algorithms are trained and models are created. In this proposal, we have taken 6335 news with equal proportionality of fake and real. Data is saved in Comma Separated Value format. This data set is divided in the ratio of 80:20 for training and testing of algorithms.

2 .Data Cleaning

Data set is a chunk of data that is in raw form. It may contain certain symbols like digits, special characters, blank lines, and data without any label. These symbols should be removed as it is of no importance and it can also affect the performance of the model in an adverse manner.

3. Data Preprocessing

After data cleaning, data is now free of unwanted symbols. This data should be converted into the form which can be used for extracting the features easily. It includes the following process:

•Converting each of the word in lowercase to avoid ambiguity between cases.
•Removal of the words which contain just one letter.
•Removal of the words that contain digits.
•Tokenize the data and removal of punctuations
•Removal of empty tokens.
•Stop Words: Stop words are useless words for NLP like “the”, “a”, “an”, “in”. This should be removed.
•Stemming:process of slicing the end or beginning of words with the intension of removing affixes.

4. Feature Selection

Feature Selection is the process where we select those features which contribute most to your prediction variable or output. Certain features are present in fake news, we have to extract them and accordingly, our classier is trained to predict the news. Here feature are the important words which appear in news. Natural language processing methods which are used to extract features are:

1)Bag of Words:

In this, each document is converted into a vector. First, all the words are taken out of the dataset forming bag of words. The vector of each document is the frequency of words present in it out of bag of word.

2) TF-IDF:

Bag of word depends only on the frequency of the words only. Certain words have very high frequency but are of no importance. To avoid this inverse document frequency is added which downscales words importance that appears a lot across documents. TF-IDF is word frequency scores that try to highlight more interesting words.

Here we use TF-IDF.

DEPLOYMENT

As CSM with TF-IDF came out to be the best method for the prediction, we need to deploy this model by making it platform independent and creating a user interface. Development of user interface and deployment of the model involves following steps:

1 .Pickling

Pickle is used for serializing and de-serializing Python object structures. Serialization refers to the process of converting an object in memory to a byte stream that can be stored on disk. The trained model is saved to disk using pickling by which we need not train the model every time we need it. We just need to deserialize it for predicting news input.

2. Pipeline

The data entered by the user will be in raw form. It should go through various intermediate processes like cleaning, processing, etc. before it can be used for prediction. Pipeline decides the flow of the data i.e. in what sequence data will go through various processes.

3. User Interface

An interface is created user to access the model without going into background detail of processing. It also removes the platform dependence i.e. it will only run on software on which it is modeled. In this project, a website constructed by HTML and CSS is used as a user interface.

4.API

API is the tool that is used to create an interface between the two applications. As our prediction model and user interface i.e. website are two applications that should be linked with each other to predict the input news and display it on the webpage.

DELIMITATION

1.Our system does not guarentee 100% accuracy.
2.The system is unable to test data that is unrelated to the training dataset.

CONCLUSION

A platform i.e. website is created which is linked to a trained Machine Learning model. CSM with TF-IDF NLP method is trained and use for the prediction of new news input by the user. This trained model link to the platform is capable of predicting news to be fake or real with an accuracy of 93%. In this platform, users can enter the news and it will predict whether the news is real or fake. It will show output as real or fake with the possibility of news to be real or fake.

FUTURE ENHANCEMENT

There is no platform via which the user can check the credibility and authenticity of the news and where authorities can directly inform about the fake news prevailing. Due to which people can believe in the news which can be a trouble for them and as well for society also. In the existing system, the action is taken after the adverse impact had already hit society. The proposed platform is useful for both common people and official authorities to prevent the spread of rumours in form of news.

NATURAL LANGUAGE PROCESSING(NLP)

As human beings,when we read a sentense or a paragraph,we can interpret the words with the whole document and understand the context.

Given today’s volume of news ,it is possible to teach to a computer how to read and understand the differences between real news and the fake news using Natural Language Processing(NLP).

The building blocks are Data set and Machine Learning Algorithms.

COSINE SIMILARITY ALGORITHM

Cosine similarity is a metric used to determine how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. In this context, the two vectors I am talking about are arrays containing the word counts of two documents. When plotted on a multi-dimensional space, where each dimension corresponds to a word in the document, the cosine similarity captures the orientation (the angle) of the documents and not the magnitude. If you want the magnitude, compute the Euclidean distance instead. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance because of the size they could still have a smaller angle between them. Smaller the angle, higher the similarity.

CLASSIFIERS

In our model, we used 1 type of machine learning algorithms and for the implementation work, we used the Jupyter notebook platform with the assistance of Python programmable language. The classification models that we implemented using the above-mentioned dataset are the passive-aggressive classifier. This algorithms are good for various classifications and that they got their properties and performance supported different datasets. For analysis and classification problems we have used four features including id or URL, title or headline, text or body, label or target. Features like date, subject are eliminated.The frequent words used in the news content are expressed using a word cloud.

Passive Aggressive Classifiers

The Passive-Aggressive algorithms are a family of Machine learning algorithms that aren't alright known by beginners and even intermediate Machine Learning enthusiasts. However, they will be very useful and efficient surely applications . Passive-Aggressive is considered algorithms that perform online learning.Their characteristics is that they remain passive when dealing with an outcome that has been correctly classified and become aggressive when a miscalculation takes place,thus constantly self-updating and adjusting.

DEVELOPING ENVIRONMENT:

- Hardware specification:

  Processor : Intel Pentium Core i3 and above

  Primary Memory : 4 GB RAM and above

  Storage : 500 GB hard disk and above

- Software specification:

  Language :Python

  Front end : Python Django

  Back end : SQLite

  Operating system : windows 7 and above

  IDE : Visual Studio code,Jupyter Notebook

  Others : HTML,CSS

  Algorithm:Cosine Similarity Algorithm-TF-IDF vectorizer

  Technique:Natural Language Processing

  Data set:Fake news & True news from Kaggle website

PROJECT PLAN:

User Story ID	Task Name	Start Date	End Date	Days	Status
1	Sprint 1	27/12/2021	27/12/2021	2	Completed
2		28/12/2021	28/12/2021		Completed
3	Sprint 2	29/12/2021	29/12/2021	5	Completed
4		15/01/2022	16/01/2022		Completed
5		22/01/2022	22/01/2022		Completed
6		23/01/2022	23/01/2022		Completed
7	Sprint 3	26/01/2022	29/01/2022	3	Completed
8	Sprint 4	05/02/2022	06/02/2022	3	Completed
9		12/02/2022	12/02/2022		Completed

USER STORY:

User Story ID	As a <type of user>	I want to <perform some task>	So that I can <achieve some goal>
1	User	Collection of Dataset	Fake news Dataset
2	User	Preprocessing of collected data(null values elimination)	Cleaned final dataset
3	User	Visualisation of input data	Graphical representation of data
4	User	Count no. of fake news & true news	Number of datas
5	User	Split data into training & testing set	80% -training data & 20%-testing
6	User	Train the data using TF-IDF vectorizer	Trained data
7	User	UI designing(a form to enter news for checking)	A form to enter news
8	User	Collect input from user & Testing	input from user & Tested data
9	User	generate output	Result(using Cosine similarity algorithm)

PRODUCT BACKLOG:

User Story ID	Priority<High /Medium/Low >	Size(Hours)	Sprint	Status<Planned/Inprogress/ Completed>	Release Date	Release Goal
1	Medium	2	1	Completed	27/12/2021	Collection of datasets
2	High	3		Completed	28/12/2021	Preprocessing of collected data
3	Medium	3	2	Completed	29/12/2021	Visualisation of input data
4	High	2		Completed	15/01/2022-16/01/2022	Count no. of fake news & true news
5	Medium	5		Completed	22/01/2022	Split data into training & testing set
6	High	5		Completed	23/01/2022	Train the data
7	High	10	3	Completed	26/01/2022, 29/01/2022	UI designing(a form to enter news for checking)
8	High	20	4	Completed	05/02/2022-06/02/2022	Testing & collect input from user
9	High			Completed	12/02/2022	Generate Result

SPRINT PLAN:

Backlog item	Status & Completion date	Original Estimate in hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13
User Story#1,2		Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours
Data collection	27/12/2021	2	2	0	0	0	0	0	0	0	0	0	0	0	0
Preprocessing	28/12/2021	3	0	3	0	0	0	0	0	0	0	0	0	0	0
User story #3,4,5,6															
Visualisation & Training	23/01/2022	15	0	0	3	3	3	3	3	0	0	0	0	0	0
User story #7															
UI Designing	29/01/2022	10	0	0	0	0	0	0	0	4	4	2	0	0	0
User story #8,9															
Testing	12/02/2022	20	0	0	0	0	0	0	0	0	0	0	8	6	6
Total		50	2	3	3	3	3	3	3	4	4	2	8	6	6

Sprint1 Actual:

Backlog item	**Status & Completion date**	**Original Estimate in hours**	**Day 1**	**Day 2**	**Day 3**	**Day 4**	**Day 5**	**Day 6**	**Day 7**	**Day 8**	**Day 9**	**Day 10**	**Day 11**	**Day 12**	**Day 13**	**Completed<Y/N>**
User Story#1,2		Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours
Data collection	27/12/2021	2	2	0	0	0	0	0	0	0	0	0	0	0	0	Y
Preprocessing	28/12/2021	3	0	3	0	0	0	0	0	0	0	0	0	0	0	Y
User story #3,4,5,6																
Visualisation & Training	23/01/2022	15	0	0	3	3	3	3	3	0	0	0	0	0	0	N
User story #7																
UI Designing	29/01/2022	10	0	0	0	0	0	0	0	4	4	2	0	0	0	N
User story #8,9																
Testing	12/02/2022	20	0	0	0	0	0	0	0	0	0	0	8	6	6	N
Total		50	2	3	3	3	3	3	3	4	4	2	8	6	6	N

Sprint2 Actual:

Backlog item	Status & Completion date	Original Estimate in hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Completed<Y/N>
User Story#1,2		Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours
Data collection	27/12/2021	2	2	0	0	0	0	0	0	0	0	0	0	0	0	Y
Preprocessing	28/12/2021	3	0	3	0	0	0	0	0	0	0	0	0	0	0	Y
User story #3,4,5,6																
Visualisation & Training	23/01/2022	15	0	0	3	3	3	3	3	0	0	0	0	0	0	Y
User story #7																
UI Designing	29/01/2022	10	0	0	0	0	0	0	0	4	4	2	0	0	0	N
User story #8,9																
Testing	12/02/2022	20	0	0	0	0	0	0	0	0	0	0	8	6	6	N
Total		50	2	3	3	3	3	3	3	4	4	2	8	6	6	N

Sprint3 Actual:

Backlog item	Status & Completion date	Original Estimate in hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Completed<Y/N>
User Story#1,2		Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours
Data collection	27/12/2021	2	2	0	0	0	0	0	0	0	0	0	0	0	0	Y
Preprocessing	28/12/2021	3	0	3	0	0	0	0	0	0	0	0	0	0	0	Y
User story #3,4,5,6																
Visualisation & Training	23/01/2022	15	0	0	3	3	3	3	3	0	0	0	0	0	0	Y
User story #7																
UI Designing	29/01/2022	10	0	0	0	0	0	0	0	4	4	2	0	0	0	Y
User story #8,9																
Testing	12/02/2022	20	0	0	0	0	0	0	0	0	0	0	8	6	6	N
Total		50	2	3	3	3	3	3	3	4	4	2	8	6	6	N

Sprint4 Actual:

Backlog item	Status & Completion date	Original Estimate in hours	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Completed<Y/N>
User Story#1,2		Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours	Hours
Data collection	27/12/2021	2	2	0	0	0	0	0	0	0	0	0	0	0	0	Y
Preprocessing	28/12/2021	3	0	3	0	0	0	0	0	0	0	0	0	0	0	Y
User story #3,4,5,6																
Visualisation & Training	23/01/2022	15	0	0	3	3	3	3	3	0	0	0	0	0	0	Y
User story #7																
UI Designing	29/01/2022	10	0	0	0	0	0	0	0	4	4	2	0	0	0	Y
User story #8,9																
Testing	12/02/2022	20	0	0	0	0	0	0	0	0	0	0	8	6	6	Y
Total		50	2	3	3	3	3	3	3	4	4	2	8	6	6	Y

Thank you